INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

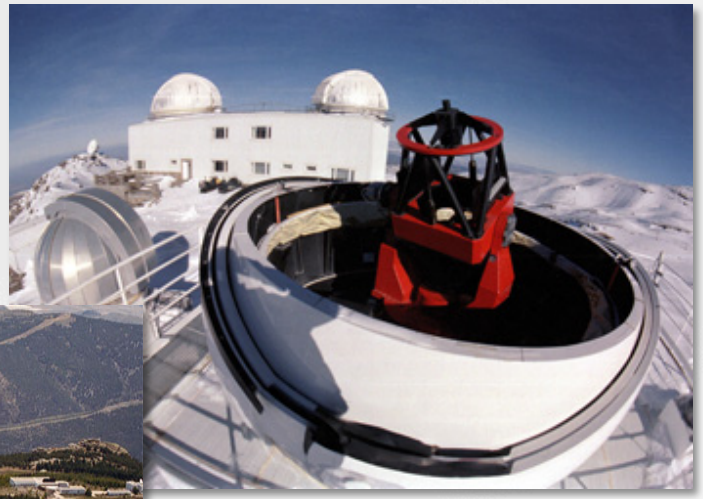
Reproducibility
The IAA-CSIC Contribution

CTA Data Model f2f Meeting

Madrid - September 20th 2016

José Enrique Ruiz
IAA - CSIC

IAA – CSIC




0 100 200 km
0 100 200 mi
Bay of Biscay
FRANCE
ANDORRA
Pyrenees
PORT.
MADRID
Balearic Sea
Balearic Islands
Granada
Mediterranean Sea
Strait of Gibraltar
North Atlantic Ocean
MOROCCO
ALGERIA
Canary Islands are not shown.

IAA – CSIC

Iván Agudo - **Coordinator/responsible of the IAA-CSIC Group**

Millimeter and radio polarimetric observations (also VLBI), optical polarization, and multi-spectral-range observations. Blazars, relativistic jets in active galactic nuclei, and the surrounding of the supermassive black hole in Sagittarius A*.

Antxon Alberdi

Radioastronomy, Radio Interferometrometry. Starbursts, AGN Relativistic Jets, Radio Supernovae, Supernova Remnants, NIR-Interferometry studies of Massive Stars.

Alberto Castro-Tirado

Optical astronomy and astrophysics with robotic telescopes. PI of the BOOTES Network of robotic telescopes, and scientific contributor to the GRANAT mission and ESA's International Gamma-Ray Laboratory INTEGRAL. Microquasars, GRBs and transients. Astrophysics from multi-spectral-range data-sets, also involving gamma-rays, but with a main focus on optical observations.

José Luis Gómez

Ultra high resolution radio and millimeter very long baseline interferometry, including VLBI with orbiting antennas. Study of AGN jets through VLBI and multiwavelength observations and their comparison with numerical simulations.

Miguel Angel Pérez-Torres

Radio astrophysics, radio interferometry. Starbursts, AGNs and supernovae and [illegible], using multi-wavelength observations, with an emphasis in radio observati[illegible]

José Enrique Ruiz

Technical aspects of astronomical Archives, Virtual Observatory and [illegible]ows[illegible]. Capture and preservation of the scientific process in order to increase reproduc[illegible]nity.[illegible]s and reuse of methodology used in the process of analysis and research in astronomy..

CTA-Spain
April 11th 2016

CTA-Consortium
August 30th 2016

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

Technical Experience

VO Archives - Modelling and Implementation

- B0DEGA – Interferometric DataCubes of Galaxies
- TAPAS IRAM 30m – Submm. Single-dish Observations
- DSS-63 Robledo 70m – Radio Single-dish Observations
- AMIGA Catalog – Physical Properties of Galaxies

IVOA Contributions

- Note. Scientific Workflows in the VO
- REC. PDL Parameter Description Language
- Draft. N-Dimensional Cube Model

Software Development

- AstroTaverna – Building workflows in the VO
- GUIPSY– Kinematic modelling for velocity datacubes of galaxies

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

The Wf4Ever Project

Wf4Ever - Advanced Workflow Preservation Technologies for Enhanced Science
2011-2013 EU FP7

1. Intelligent Software Components (ISOC
2. University of Manchester (UNIMA
3. Universidad Politécnica de M , Spain)
4. Poznan Supercomputing king Centre (PSNC, Poland)
5. Universisty of Oxford n Centre (OXF, UK)
6. Instituto Astrofísi a (IAA-CSIC, Spain)
7. Leiden Unive al Centre (LUMC, NL)

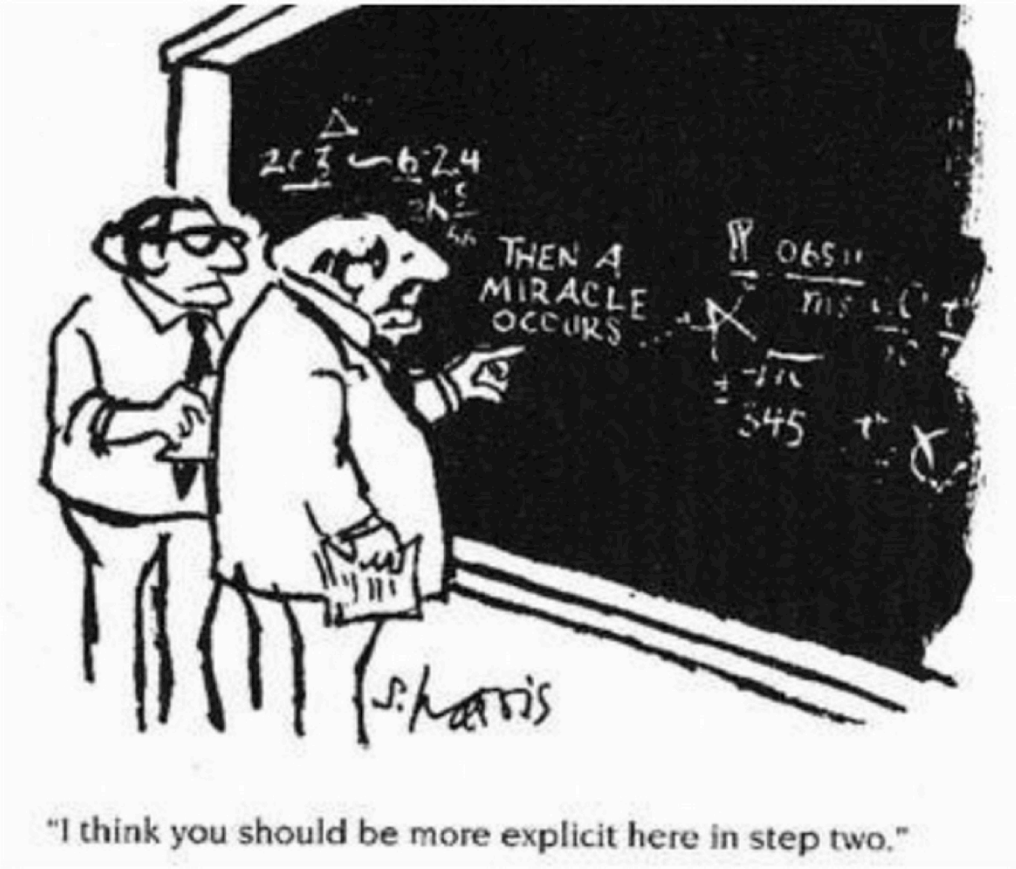
Reproducible Science

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

The Reproducibility Crisis

"... up to 70% of research from academic labs **cannot be reproduced**, representing an enormous waste of money and effort."

- Elizabeth Iorns, Science Exchange



"I think you should be more explicit here in step two."

Reproc

Announcement: Reducing our irreproducibility : Nature News & Comment

www.nature.com/news/announcement-reduci

nature.com : Sitemap Login : Register

nature International weekly journal of science

Search Go Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 496 > Issue 7446 > Editorial > Article

NATURE | EDITORIAL

Announcemer

24 April 2013

PDF | Rights & Perm

Over the past year, *Nature* h

reliability and reproducibility

Science AAAS.OR

AAAS NEWS

Science The World's Lea

Science Home | Current Issue | P

Home > *Science* Magazine > 17 January

Article Views
- Summary
- Full Text
- Full Text (PDF)

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Rights & Permissions
- Commercial Reprints and E-Prints
- View PubMed Citation

Related Content

Science 1
Vol. 343
DOI: 10.1

EDITORI

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Science*.

Science advances on a foundation of trusted discoveries. Reproducing an experiment i
approach that scientists use to gain confidence in their conclusions. Recently, the scie
community was shaken by reports that a troubling proportion of peer-reviewed precli
not reproducible. Because confidence in results is of paramount importance to the bro
community, we are announcing new initiatives to increase confidence in the studies pu
Science. For preclinical studies (one of the targets of recent concern), we will be adopt
recommendations of the U.S. National Institute of Neurological Disorders and Stroke (I
increasing transparency.* Authors will indicate whether there was a pre-experimental
handling (such as how to deal with outliers), whether they conducted a sample size es
ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and
experimenter was blind to the conduct of the experiment. These criteria will be included in our author
guidelines.

The Economist

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

OCTOBER 19TH–25TH 2013 Economist.com

HOW
SCIENCE
GOES
WRONG.
99
Einsteinium

Los Angeles Times | BUSINESS

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH STYLE TRAV

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true.

Email | Share 9K | Tweet 1,076 | Like 7.5k | LinkedIn 85 | +1 299

By Michael Hiltzik
October 27, 2013

In today's world, brimful as it is with opinion and falsehoods masquerading as facts, you'd think the one place you can depend on for verifiable facts is science.

You'd be wrong. Many billions of dollars' worth of wrong.

journal of science

anced search Search Go

article

NATURE | EDITORIAL

Must try harder

Nature **483**, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

PDF | Citation | Reprints | Rights & permissions | Article metrics

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Digital Astronomy

Astronomy research lifecycle is **entirely digital**

- Observation proposals
- Data reduction pipelines

- Analysis of science ready data
- Catalogs of objects and data archives
- Publish process
  - Final data results
  - Experiment in Digital Libraries ADS/arXiv

Reproducible research is still not possible in a digital world

A rich infrastructure of data is not efficiently used

A normalized preservation of methodology is needed

Tools

Visibility, Efficiency and Re-use

Optimize return on investments for big facilities

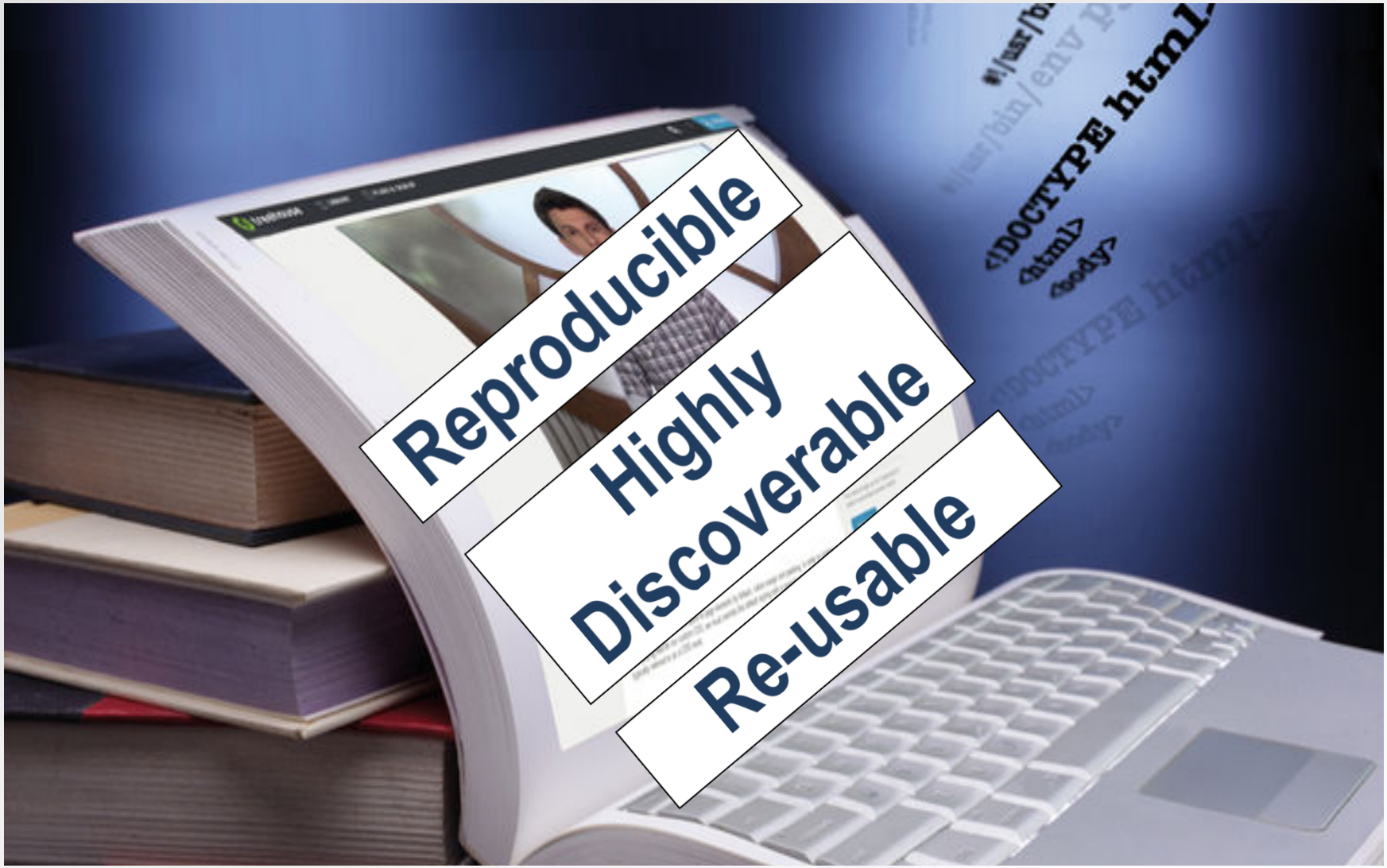
- Avoid duplication of efforts and reinvention
- How to discover and not duplicate ?
- How to re-use and not duplicate ?
- How to make use of best practices ?
- How to use the rich infrastructure of data ?
- **Intellectual contributions encoded in software**

More data in archives do not imply more knowledge

- Expose **complete scientific record**, not the story
- Allow easy **discovery** of methods and tools

The Executable Paper

Time has come to go **beyond the PDF**



Barriers

Barriers to Data and Code Sharing in Computational Science

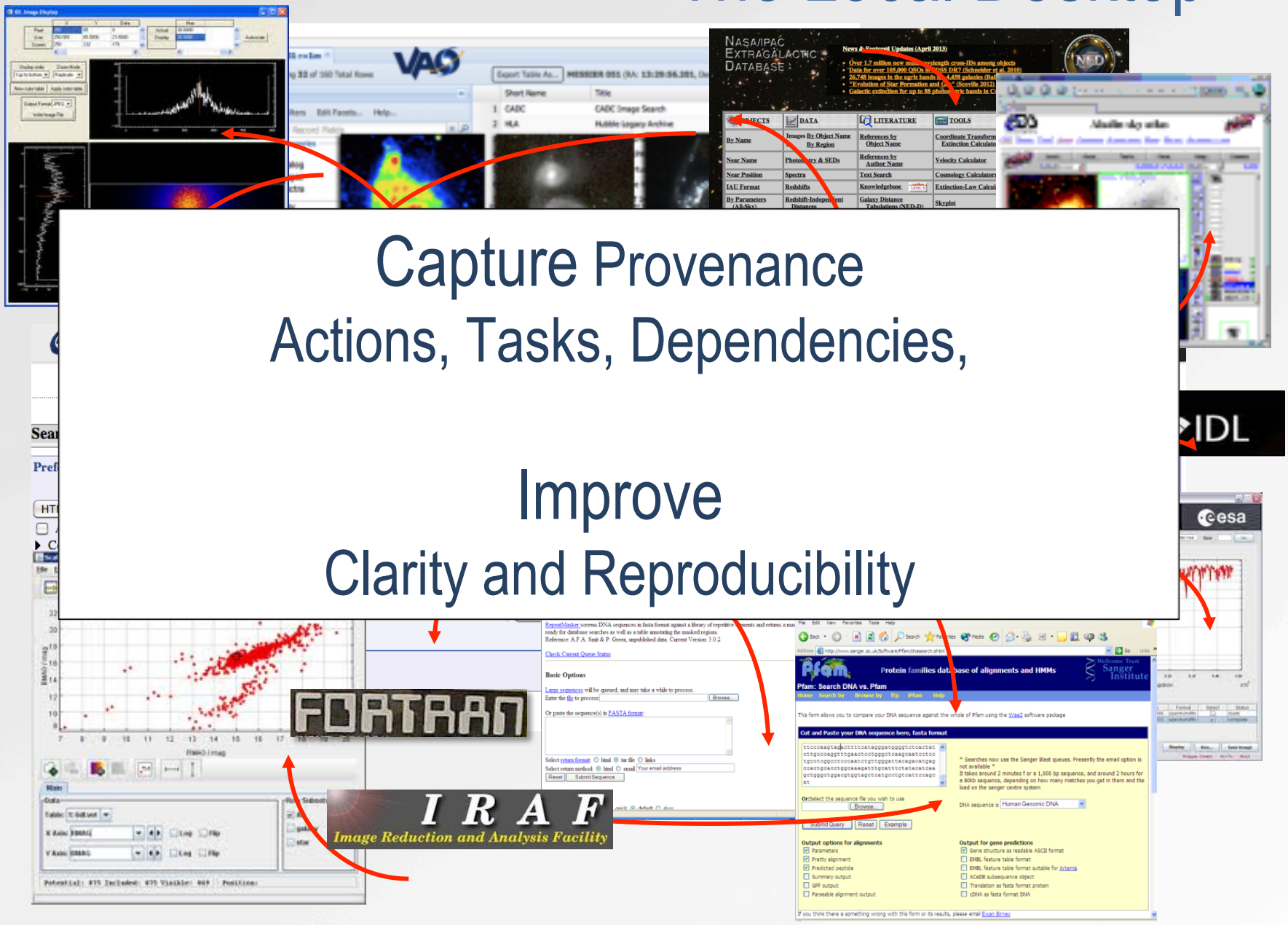
Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	

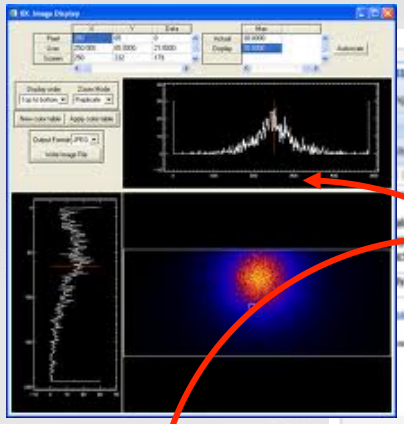
Incentives

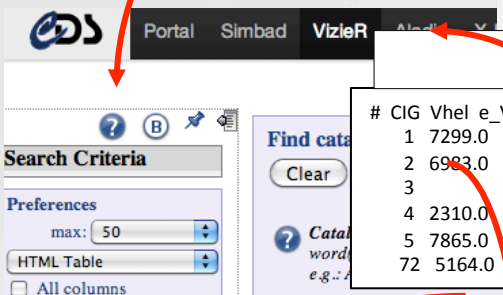
Tools

The Local Desktop

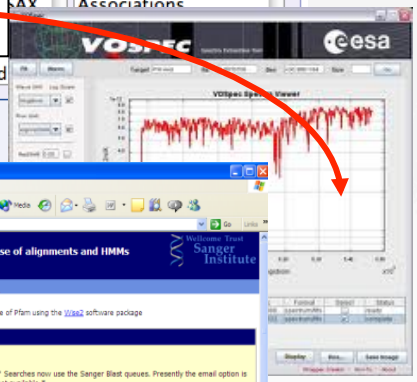


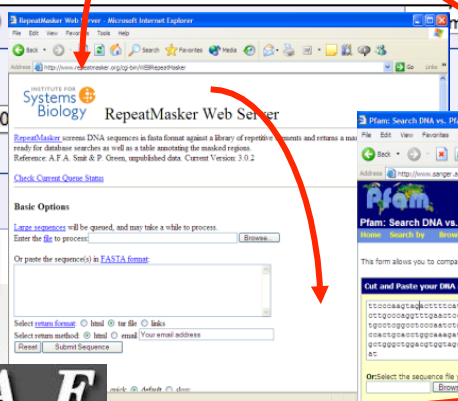
The Local Desktop

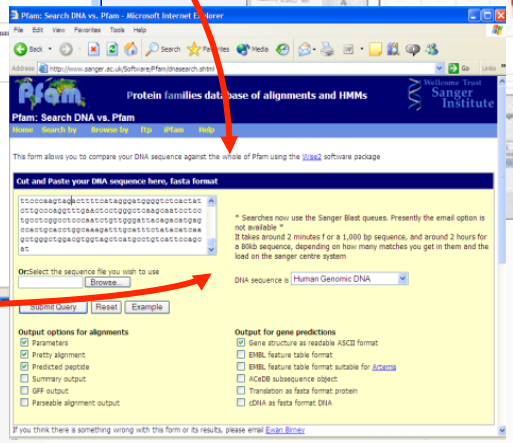




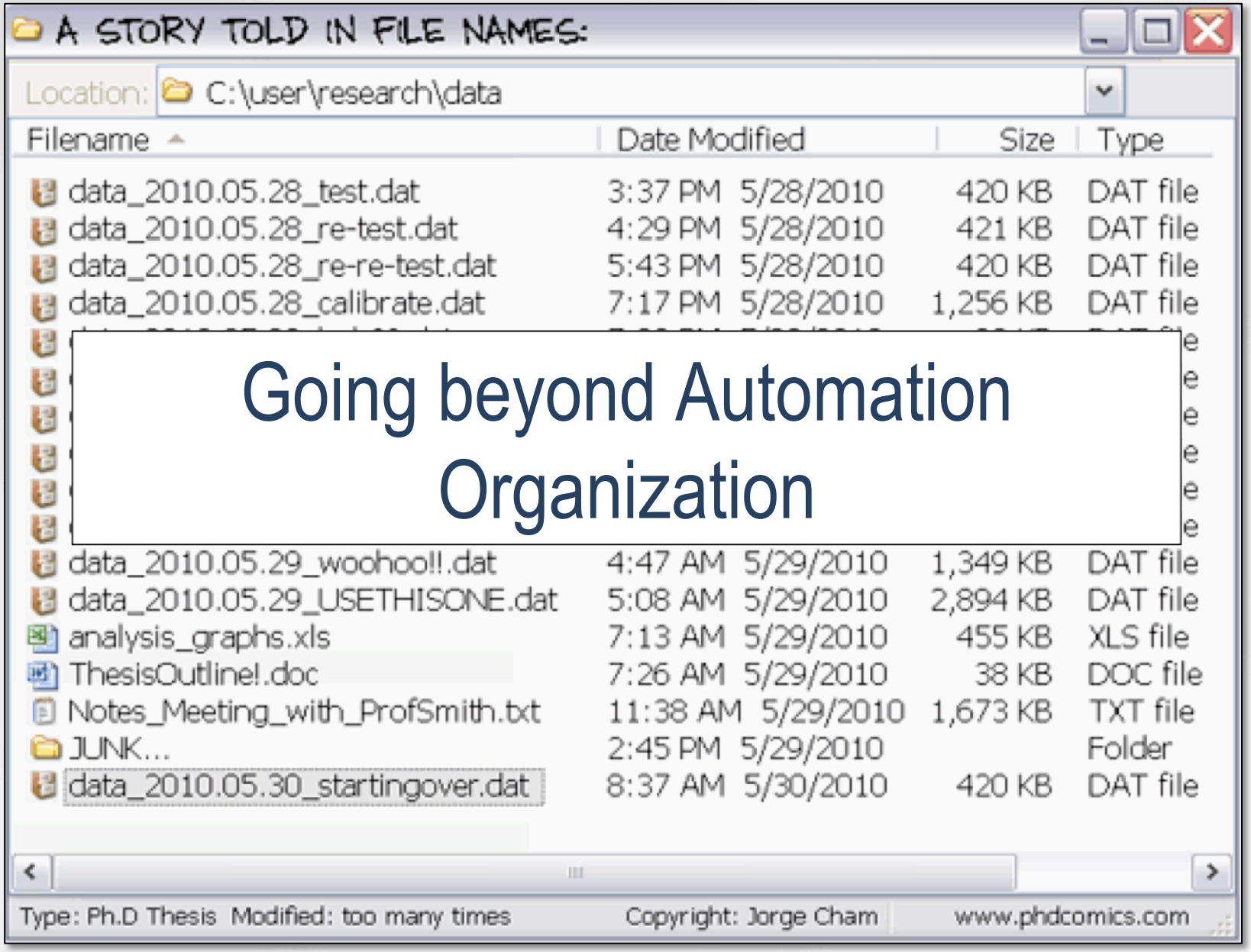

```
# CIG  Vhel  e_Vhel  r_Vhel  Dist  MType  e_MType  OptAssym  r_MType  Bmag  e_Bmag
  1  7299.0   3.0   1   96.9   5.0   1.5   1   1   14.167   0.271   0.173  0.571  0.040  13.383
  2  6983.0   6.0   2   94.7   6.0   1.5   0   1   15.722   0.324   0.255  0.278  0.031  15.157
  3                        4.0   1.5   0   1   16.057   0.507   0.246  0.354         15.457
  4  2310.0   1.0   3   31.9   3.0   1.5   0   1   12.818   0.424   0.252  0.863  0.017  11.685
  5  7865.0  10.0   3  105.9   0.0   1.5   0   1   15.602   0.364   0.225  0.131  0.118  15.128
 72  5164.0   9.0   2   68.5   5.0   1.5   1   1   14.445   0.325   0.315  0.367  0.028  13.735
```







The Local Desktop

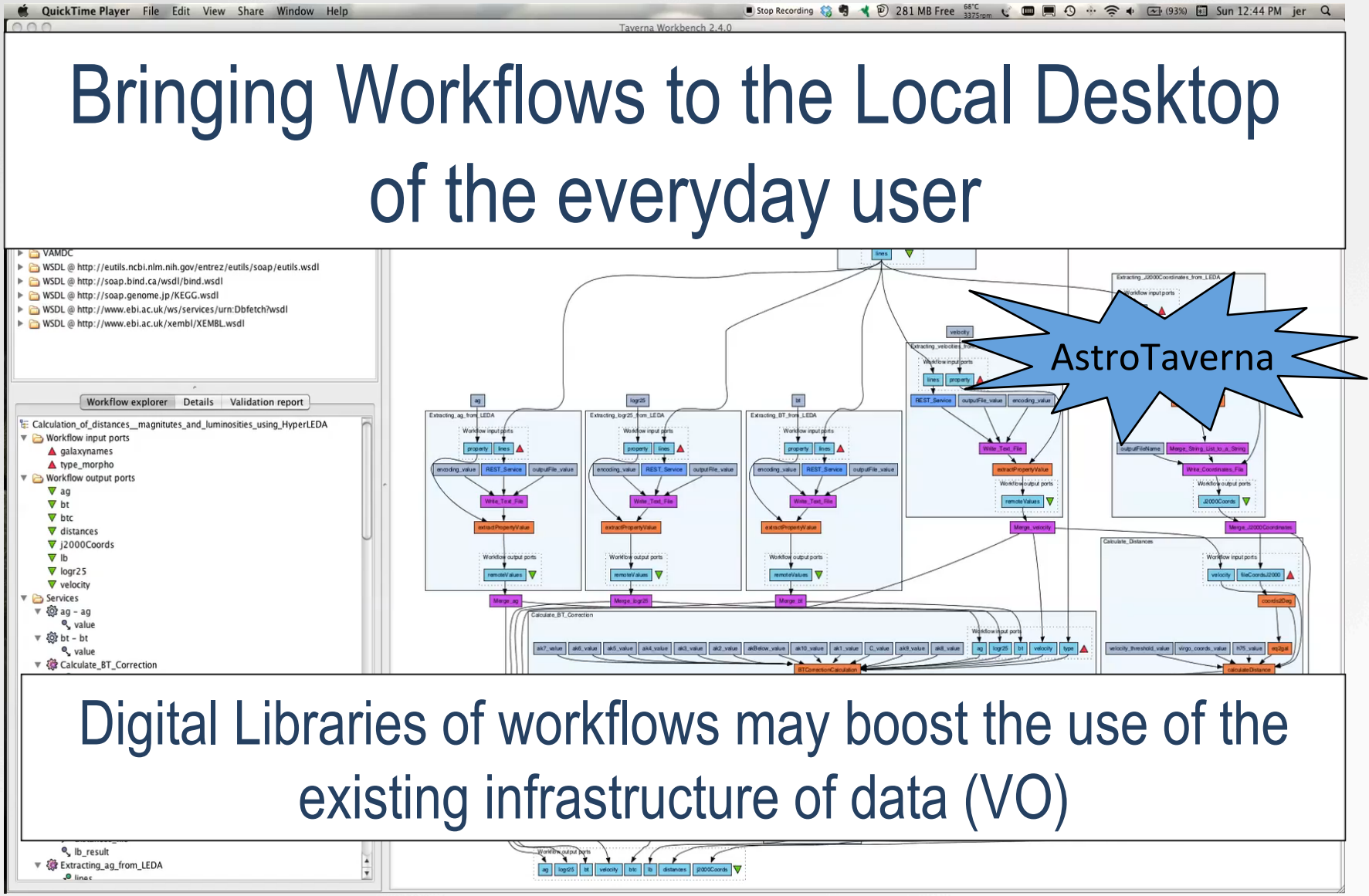


Definitions and Requirements

- Understandable
- Inspectable /Browseable
- Reproducible
  - Reproduce in a different environment
- Repeatable
  - Obtain same results when reproduced
- Re-usable
  - Execute with different data
- Re-purposable /Modular /Extensible
  - Modify for a different purpose
- Discoverable
- Socially Curated
  - Annotated, Recommended, Rated

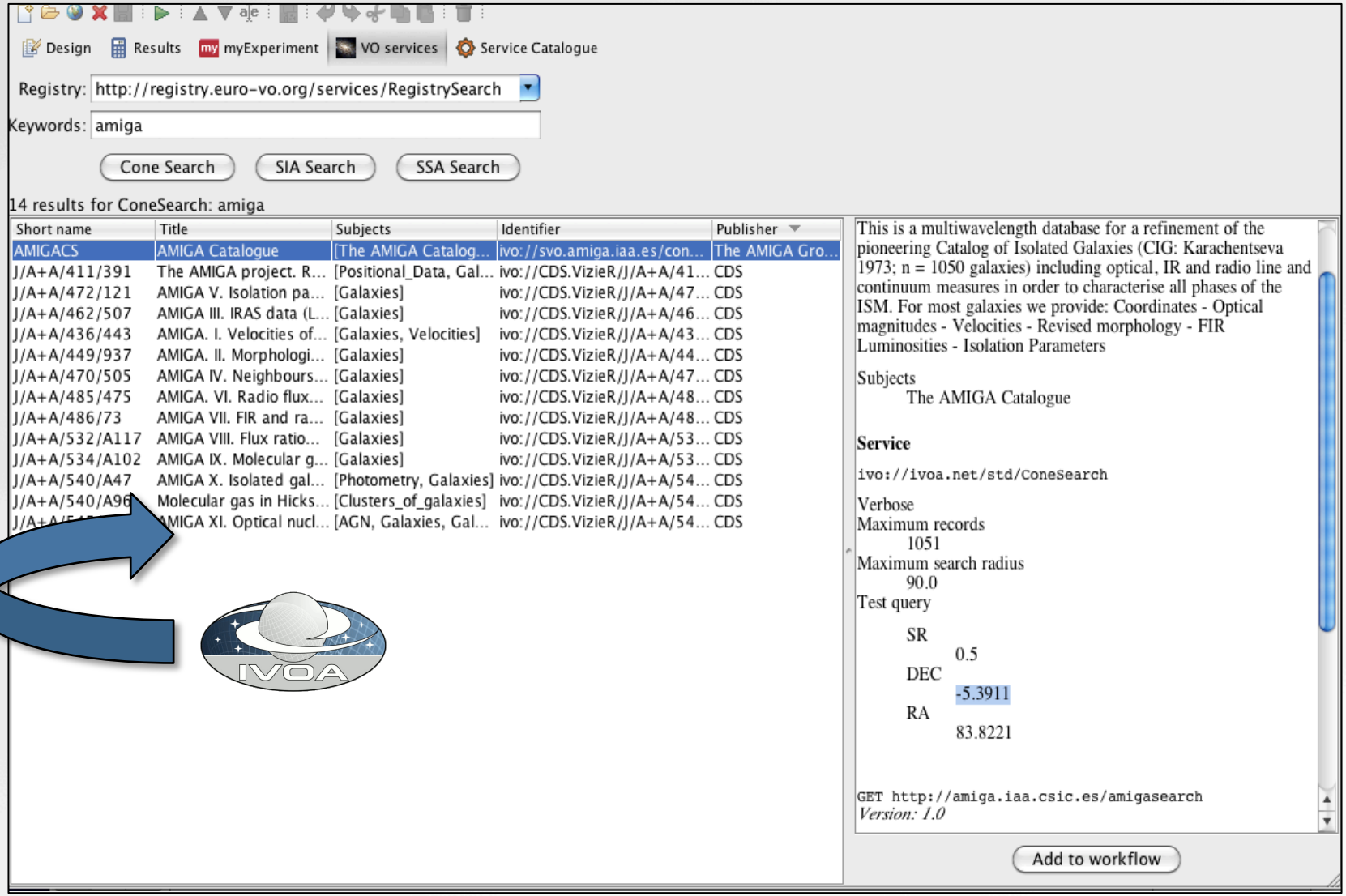
Scientific Workflows

Bringing Workflows to the Local Desktop of the everyday user

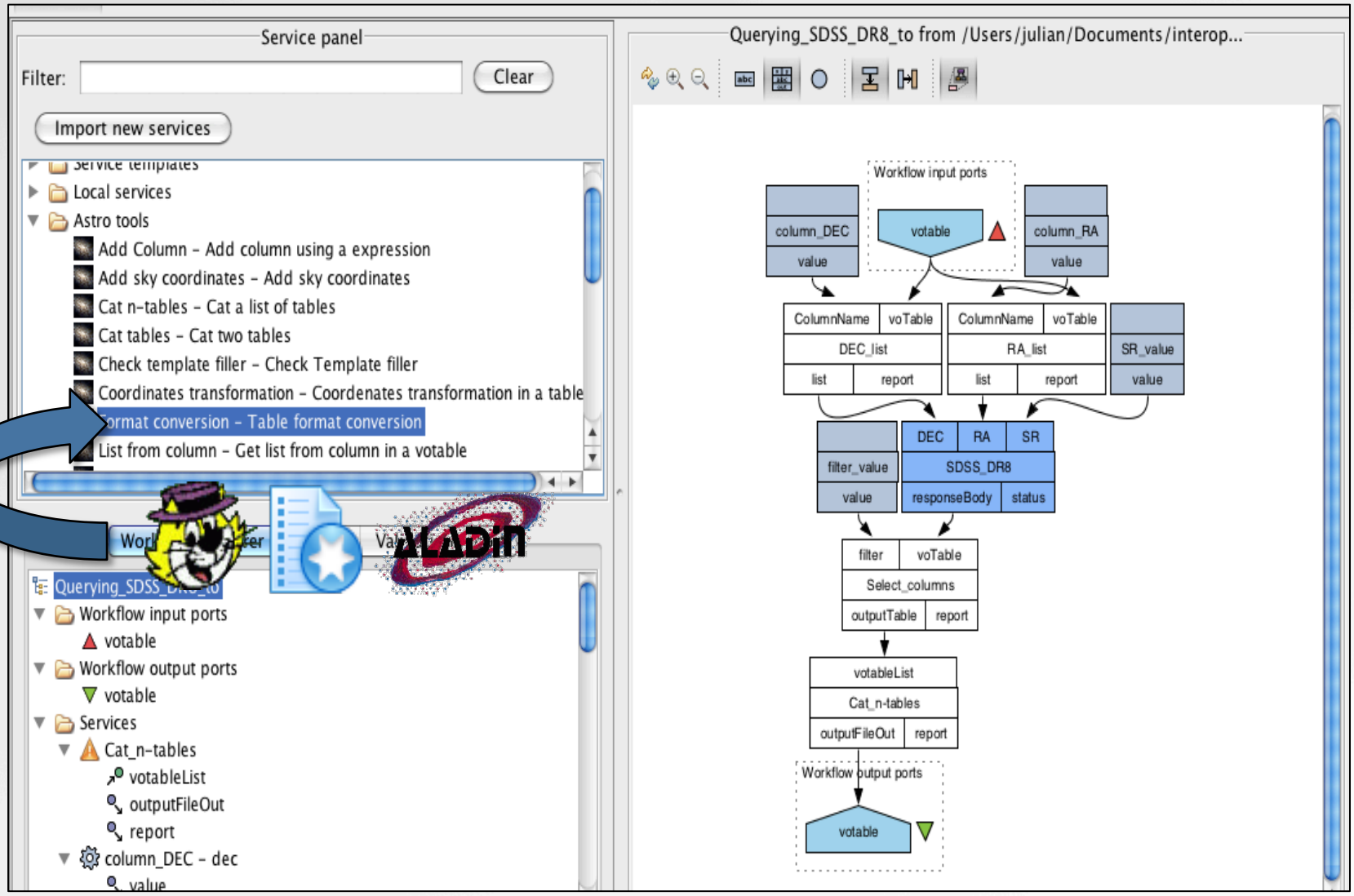


Digital Libraries of workflows may boost the use of the existing infrastructure of data (VO)

AstroTaverna

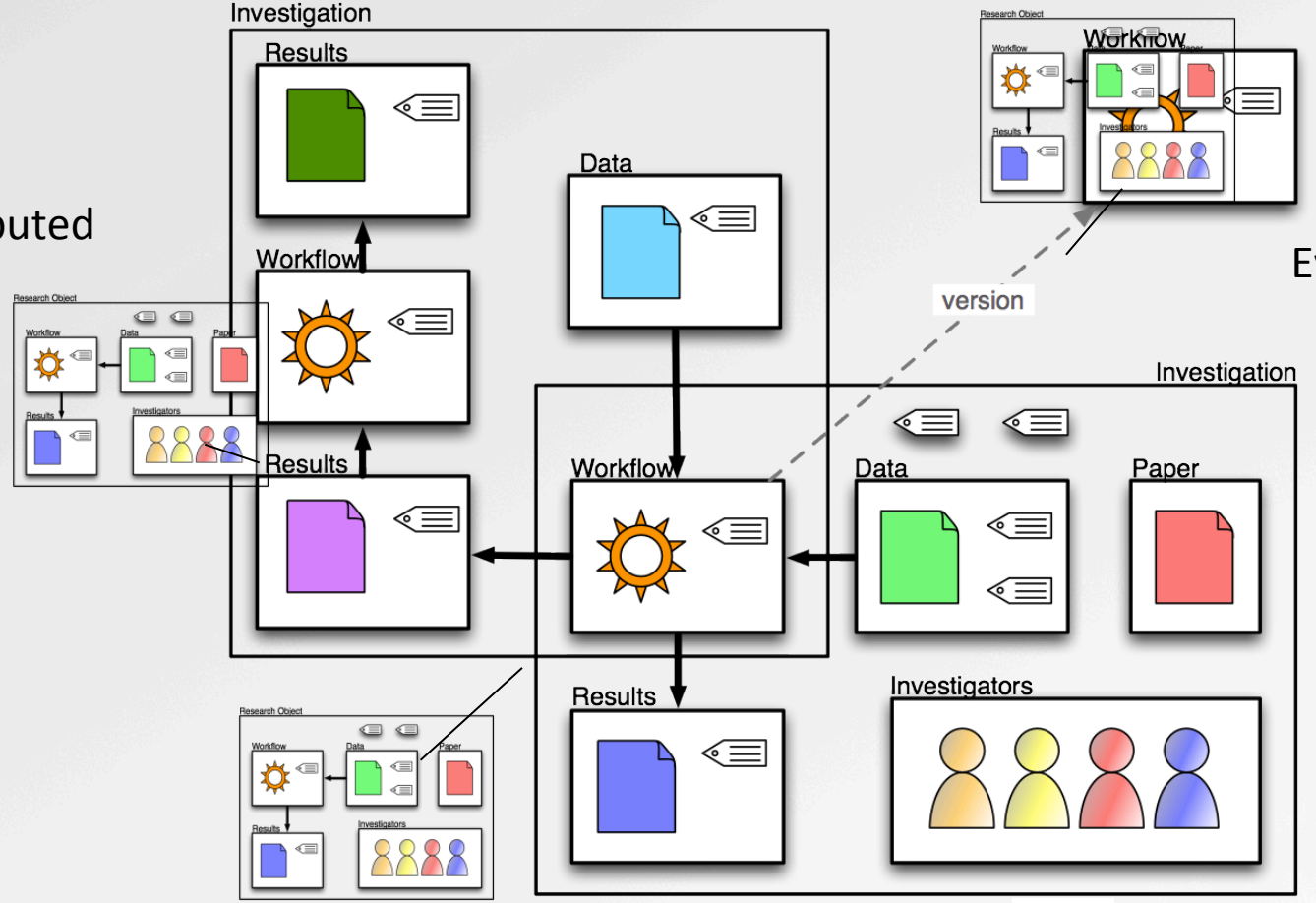


AstroTaverna



Workflow-Centric Research Objects

Modular distributed aggregation of digital resources



Distributed

Evolving

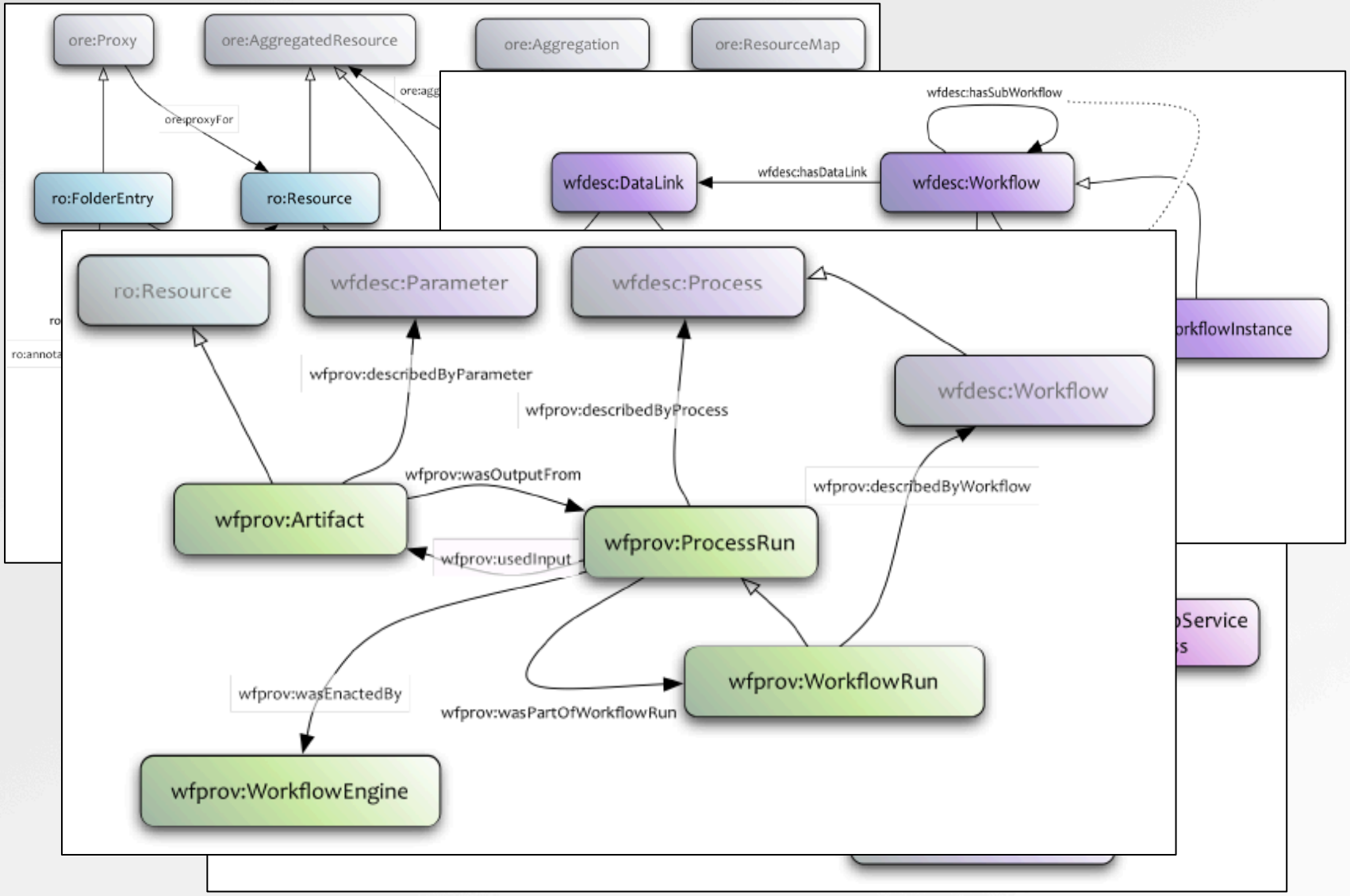
Technical Objects

Social Objects

Workflow-Centric Research Objects

- Ontologies and Vocabularies
  - OAI Objects Reuse and Exchange
  - Open Annotation Ontology
  - PROV Ontology
  - WfProv Ontology
  - WfDesc Ontology
  - RO Ontology
- RO Models
  - Aggregation
  - Provenance
  - Evolution
  - Annotations
  - Checklists (Minimum Information Model)
  - Social Recommendations
- RO Bundle Packaging

RO Models



RO RDF Serializations

```
<http://sandbox.wf4ever-project.org/rosrs5/ROs/HyperLEDALuminosities/>
    dct:created "2012-02-01T09:33:56.832Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
    dct:creator <https://www.google.com/accounts/o8/id?id=AItOawl7tEZAtdiQmIigvXcU4KkhGsBzgm0gfKM> ;
    ore:isDescribedBy <http://sandbox.wf4ever-project.org/rosrs5/ROs/HyperLEDALuminosities/.ro/manifest> ;
    a
    or
```

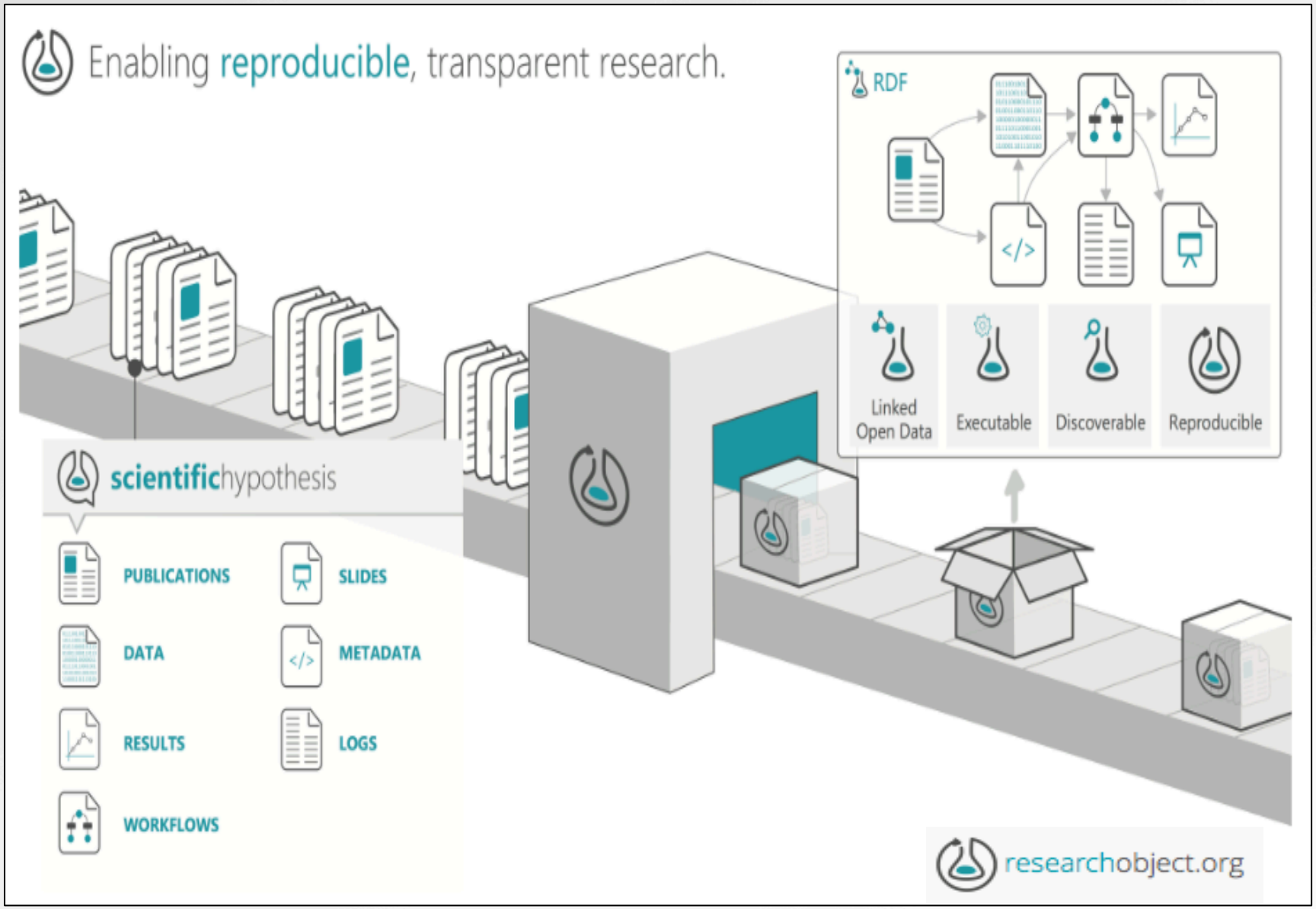
```
    <http://leda.univ-lyon1.fr/ledacat.cgi>
        a <http://purl.org/wf4ever/wf4ever#WebServiceProcessTemplate> .

    <http://sandbox.w
        wfprov:descri
        wfprov:usedIn
        a <http://pur

    <http://sandbox.w
        a <http://pur

    <http://sandbox.w
        wfprov:wasOut
        a <http://pur

    <http://sandbox.w
        wfprov:wasOut
        a <http://pur

    <htt
```

```
        <http://
            ao:b
            dct:
            dct:
            ao:a
            a ao

        <http://
            ao:b
            dct:
            dct:
            ao:a
            a ao:Annotation .

        _:A0
            a <http://xmlns.com/foaf/0.1/Agent> ;
            foaf:name "Jose Enrique Ruiz" .
```

```
    a <http://purl.org/wf4ever/wf4ever#Dataset> .
```

```
@prefix wfprov: <http://purl.org/wf4ever/wfprov#> .

  :wf1 a wfprov:WorkflowRun ;
  wfprov:describedByWorkflow :wfTempl;
  wfprov:wasEnactedBy :Taverna.
 #Additional metadata about the workflow run could be added here: start time, end time, whether it has bees successf

:proc1 a wfprov:ProcessRun ;
  wfprov:usedInput :i1 ;
  wfprov:wasPartOfWorkflowRun :wf1;
  wfprov:describedByProcess :templProcess1.

:proc2 a wfprov:ProcessRun ;
  wfprov:usedIntput :o1 ;
  wfprov:wasPartOfWorkflowRun :wf1;
  wfprov:describedByProcess :templProcess2 .

:i1 a wfprov:Artifact;
    wfprov:describedByParameter :param1.

:o1 a wfprov:Artifact;
    wfprov:wasOutputFrom :proc1;
        wfprov:describedByParameter :param2.

:o2 a wfprov:Artifact;
    wfprov:wasOutputFrom :proc2;
        wfprov:describedByparameter :param3.

#wfdesc:parameters and templates have been omitted to leave the example as simple as possible.
```

Research Object Knowledge Hub

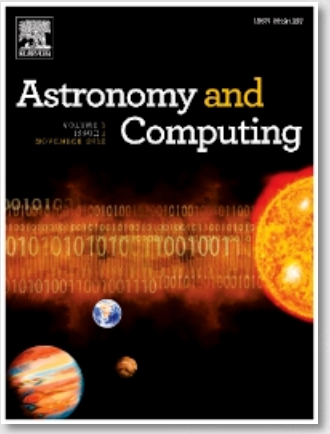


RO adoption failed

- Preservation and Discovery
  - Digital Libraries
  - Linked Data RDF
- Tools and Research Methodology
  - Tools for RO packaging/inspection/annotation missing
  - Workflow centric vs. script and data-massaging
  - Web-Services oriented workflows
  - Porting software and scripts is cumbersome
  - Non-controlled research environment
  - Steep learning curve

ELSEVIER article of the future

Related Initiatives



Graphical abstract

Source code repositories

The journal strongly encourages authors to make source code available where appropriate, especially in the case

Video data

Elsevier accepts video material and animation sequences to support and enhance your scientific research. Authors who have video or animation files that they wish to submit with their article are strongly encouraged to include links to these within the body of the article. This can be done in the same way as a figure or table by referring to the video or animation content and noting in the body text where it should be placed. All submitted files should be properly labeled so that they directly relate to the video file's content. In order to ensure that your video or animation material is directly usable, please provide the files in one of our recommended file formats with a preferred maximum size of 50 MB. Video and animation files supplied will be published online in the electronic version of your article in Elsevier Web products, including ScienceDirect: http://www.sciencedirect.com. Please supply 'stills' with your files: you can choose any frame from the video or animation or make a separate image. These will be used instead of standard icons and will personalize the link to your video data. For more detailed instructions please visit our video instruction pages at http://www.elsevier.com/artworkinstructions. Note: since video and animation cannot be embedded in the print version of the journal, please provide text for both the electronic and the print version for the

AudioSlides

MATLAB FIG files

***NEW* Inline supplementary computer code**

Elsevier now offers you the possibility to place supplementary computer code, data snippets, algorithms and other machine readable structures at the right place in your online article in reusable .txt format. This will allow readers to easily view this material in the appropriate context, and to directly copy it to the clipboard or download the original source file for testing or re-use. If you would like to have reusable "computer code" inserted into the body of your online article please indicate in your manuscript where they should be placed and number them in order of appearance, e.g. "Insert Inline Supplementary Computer Code 1 here". To support discoverability and reusability please submit these items in *.txt format and make sure to include a descriptive title and caption that references the characteristics and the appropriate environment of this material , e.g. 'An algorithm for filtering text files in R'. For more information please visit http://www.elsevier.com/ism.

Related Initiatives

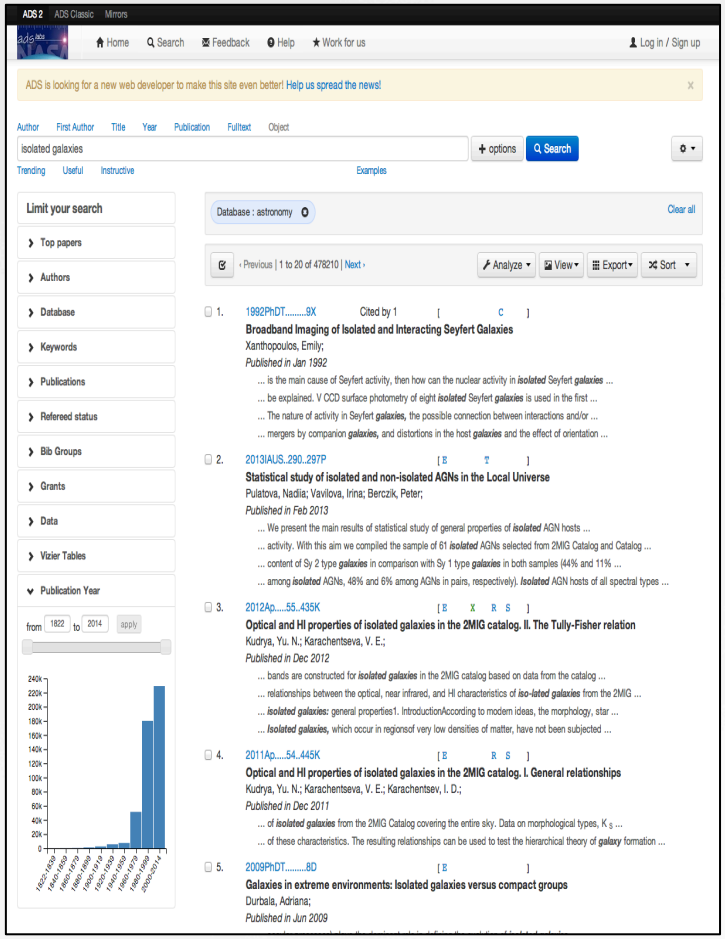
ADS Bumblebee Interface

ADO Linked Components

Faceted Browsing/Filtering

- Publications
- Journals
- Authors
- SIMBAD Objects
- Tabular data behind the plots CDS
- Observing time Proposals
- Used facilities, surveys or missions
- ASCL reference of used software
- Publication Date

https://ui.adsabs.harvard.edu/



Related Initiatives

- iPython Notebooks /JupyterLab /Binder
  - Richly documented scripts
  - Provenance of data results exposed
  - Cloud-based platforms for reproducibility
- Sumatra – Provenance capture and reproducibility
- VisTrails – Workflow and exec. provenance inspection
- ReproZip – Reproducible pack and inspection
- noWorkflow – Provenance capture in Python scripts
- Virtualization Docker /Vagrant

A Python Library for Provenance Recording and Querying

Carsten Bochner, Roland Gude, and

Simulation and Software T
German Aerospace Ce
51147 Cologne, Germ
{Carsten.Bochner,Roland.Gude,Andrea
http://www.dlr.de/

Looking Inside the Bl
Provenance using Dy

Manolis Stamatogiannakis

VU University Amst
{manolis.stamatogiannaki

Abstract. Knowing the provena
its trustworthiness. Various appr
infer data provenance. However,
ecuting program as a *black-box*,
provenance, or require developer
provenance-aware. In this paper,
proach to capturing data provena
widely used in the security and re
able to identify data provenance
tation of unmodified binaries, wit
of, their source code. Hence, we
well-known applications. Because
program, it captures high-fidelity

Keywords: data provenance, dynam

then CXXR has been regularly upgrade
of R (usually synching on the .1 minor re
CXXR has shadowed the increasing deployment of the bytecode compiler within

ES3: A Demonstration of Transparent Provenance for

Generating Scientific Documentation for Computational Experiments Using Provenance

Adianto Wibisono[1,2], Peter Bloem[1], Gerben K.D. de Vries[1], Paul Groth[2], Adam Belloum[1], Marian Bubak[1,3]

[1] System and Network Engineering Group, Informatics Institute, University of Amsterdam, The Netherlands
{a.wibisono, p.bloem, g.k.d.devries, a.z.s.belloum}@uva.nl
[2] VU University Amsterdam, The Netherlands
pgroth@vu.nl
[3] Department of Computer Science, AGH Krakow, Poland
bubak@agh.edu.pl

Abstract. Electronic notebooks are a common mechanism for scientists to document and investigate their work. With the advent of tools such as IPython Notebooks and Knitr, these notebooks allow code and data to be mixed together and published online. However, these approaches assume that all work is done in the same notebook environment. In this work, we look at generating notebook documentation from multi-environment workflows by using provenance represented in the W3C PROV model. Specifically, using PROV generated from the Ducktape workflow system, we are able to generate IPython notebooks that include results tables, provenance visualizations as well as references to the software and datasets used. The notebooks are interactive and editable, so that the user can explore and analyze the results of the experiment without re-running the workflow.
We identify specific extensions to PROV necessary for facilitating documentation generation. To evaluate, we recreate the documentation website for a paper which won the Open Science Award at the ECML/PKDD 2013 machine learning conference. We show that the documentation produced automatically by our system provides more detail and greater experimental insight than the original hand-crafted documentation. Our approach bridges the gap between user friendly notebook documentation and provenance generated by distributed heterogeneous components.

Reproducibility in CTA Research

- Computing environment
  - Observatory controlled environment
  - Astronomer desktop
- Methodology
  - Automation
  - Tools/Scripts collage
- Products
  - Observations and ancillary technical assets
  - Derived high level data products
  - Derived data/results from astronomer's desktop analysis

Reproducibility in CTA Research

- Metadata does not means central database
  - RDF Files serialization of provenance
  - SPARQL needed to query RDF TripleStore DB
  - Proper ontologies/vocabularies for CTA?
  - SQLite DB implemented in desktop by repro. tools
  - Environment captured in other formats (virtualization)
  - Digital Library of experiments?
- Provenance capture automation
  - Definition: **AST analysis** captures code
  - Deployment: `modulefinder` captures dependencies
  - Execution: **listeners+reflection** captures function calls

Reproducibility in CTA Research

- Use cases drive the design
  - Portability needs?
  - User needs for provenance information?
    - Diff based
    - Graph based
    - Query based
- Provenance in
  - Ctools?
  - Science Gateway Workflows?
  - Pipelines?
  - Monte-Carlo simulations?

Decalogue

How NOT to be a good Astronomer in XXI Century

1. In marketing just advertise your results – do not say how to reproduce them
2. Do things quickly and forget about them once you've submitted the paper
3. Be untidy – spread your code and data in a variety of formats and folders
4. Do not provide data results behind the plots– including the plots is just fine
5. Practise the "data mine-ing" – input data and/or results are mine
6. Practise the "data flirting" – please call me, if you want to know more
7. Always cite the same authors and papers or those that cite you
8. Do not reference other resources than papers – never provide URL links
9. Do not search info on Internet with other tools than ADS or arXiv
10. Do not contact others if you re-use – duplicate and reinvent for your own